

Segmentación y análisis morfológico de textos en español utilizando el sistema SMORPH¹

Salah Aït-Mokhtar - José Lázaro Rodrigo Mateos

Groupe de Recherche dans les Industries de la Langue
-GRIL-

Formation Doctorale Linguistique et Informatique
Université Blaise-Pascal (Clermont II)

34 avenue Carnot - 63037 CLERMONT-FERRAND

Tél. : 73 40 63 72 - Fax : 73 40 63 99

salah@gril.univ-bpclermont.fr

jose@gril.univ-bpclermont.fr

Abstract:

This paper describes the SMORPH system -a tool for building electronic dictionaries, morphological analysis and tagging of running texts- and how it is currently used to build a Spanish dictionary and to process Spanish texts. The system has shown to be efficient and deals, in a declarative fashion, with several typographical and morphological phenomena (punctuation and typographical rules, prefixes and suffixes, compounds, unknown words, etc.) which, though very usual in running texts, are not taken into account by most grammar building tools and parsers.

SMORPH is part of the GRIL works on the study and development of tools for automatic linguistic processing of texts. It must be then considered as a module integrated in other projects (syntactic analysis, etc.).

Keywords: text segmentation, typographical analysis, morphological analysis, text processing and tagging.

Resumen:

En este artículo presentamos el sistema SMORPH, una herramienta para la construcción de diccionarios electrónicos para el análisis morfológico y el etiquetado de textos reales, así como su utilización para la construcción de un diccionario para el tratamiento de textos en español.

El sistema se revela eficiente para declarar los fenómenos tipográficos y morfológicos (reglas tipográficas y de puntuación, prefijos y sufijos, palabras compuestas, palabras desconocidas, etc.), que pese a aparecer frecuentemente en los textos reales, no siempre se tienen suficientemente en cuenta en los analizadores sintácticos.

Smorph se enmarca en los trabajos del GRIL sobre estudio y desarrollo de herramientas de tratamiento lingüístico automático de textos, y debe considerarse como un módulo que se integra en otros proyectos (análisis sintáctico, etc.).

Palabras clave: Segmentación, análisis tipográfico, análisis morfológico, tratamiento y etiquetado de corpus.

¹ Nuestro sincero agradecimiento a Gabriel G. Bès y a Caroline Dafniet por su valiosa colaboración.

1 Presentación general de Smorph

En la figura 1 se presenta el funcionamiento general de SMORPH:

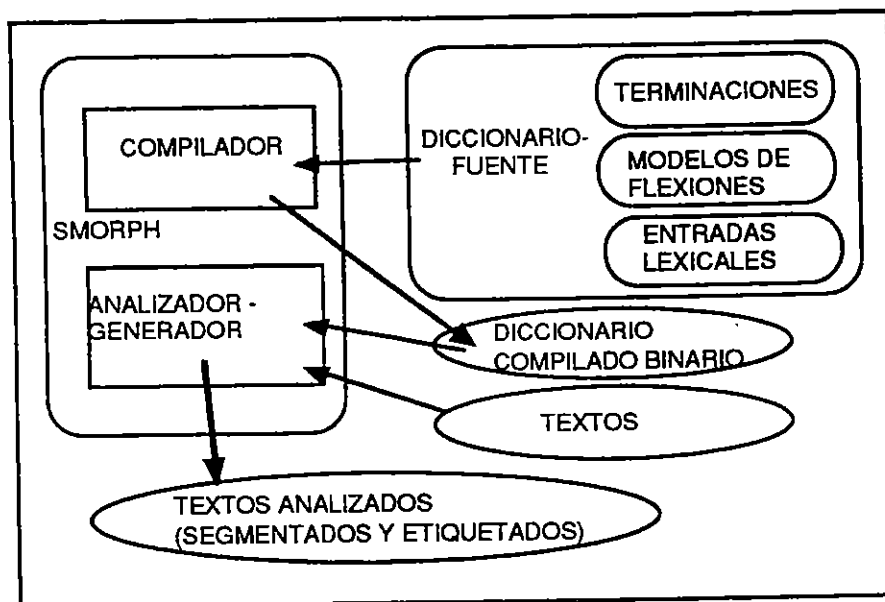


figura 1

Podemos distinguir dos partes: el sistema que está compuesto de un compilador y de un analizador generador; y los datos externos necesarios para que el sistema funcione, un diccionario fuente.

El diccionario fuente, es un conjunto formado por tres tipos de ficheros: "*terminaciones*," "*modelos de flexión morfológica*" y "*entradas lexicales*".

A partir de los datos del diccionario fuente, el compilador produce un diccionario binario, compuesto por todas las formas flexivas organizadas en forma de un autómata de estados finitos.

Este diccionario binario será utilizado por el analizador-generador de SMORPH para segmentar y etiquetar textos, y para generar formas flexivas.

2 Creación de un diccionario

Para crear un diccionario, el usuario debe definir tres tipos de datos: terminaciones, modelos de flexiones y entradas lexicales.

Un conjunto lógicamente coherente de estos datos constituye un diccionario fuente.

2.1 Terminaciones

Es necesario declarar todas las terminaciones necesarias para la definición de un modelo de flexión, en un fichero de terminaciones.

EJEMPLOS DE TERMINACIONES

#####

@ . # terminación vacía

o . # para nombres masculinos singulares

a . # para nombres femeninos singulares

os . # para nombres masculinos plurales

as . # para nombres femeninos plurales

es . # para nombres plurales terminados por consonante

terminación distinguida : para nombres femeninos singulares
ción n/sf .

terminación distinguida : para nombres femeninos plurales
ciones n/pf .

figura 2

Las terminaciones se declaran, unas seguidas de otras, separadas por puntos.

El caracter "#" permite introducir comentarios .

Es posible declarar la terminación vacía, representada con el caracter "@".

Es posible asociar a una terminación distinguida (cf. 4.4) la definición morfológica correspondiente.

2.2 Modelos de flexiones morfológicas

Una definición morfológica es una lista de valores de huellas morfológicas como por ejemplo, el nombre, persona o genero.

Un modelo de flexión reúne el conjunto de flexiones de una clase de palabras.

La descripción de esta flexión consiste en asociar a un conjunto concreto de terminaciones, un conjunto de definiciones morfológicas.

EJEMPLOS DE MODELOS NOMINALES

#####

| | | |
|-------------------|--------------|-----------------|
| # MODELO @n-os : | NOMBRES TIPO | niño -os -a -as |
| @n-os -1 | | |
| +o | n/sm | |
| +os | n/pm | |
| +a | n/sf | |
| +as | n/pf | |
| # MODELO @nm-es : | NOMBRES TIPO | árbol -es |
| @nm-es 0 | | |
| +@ | n/sm | |
| +es | n/pm | |
| #MODELO @n-s: | NOMBRES TIPO | responsable -s |
| @n-s -0 | | |
| +@ | n/s* | |
| +s | n/p* | |

figura 3

La declaración de un modelo de flexión tiene la sintaxis siguiente:

```
< nombre del modelo>  -<índice de la raíz>
                        <terminación >1 <definición morfológica>1
                        <terminación >2 <definición morfológica>2
                        .....
                        <terminación >n <definición morfológica>n
```

El primer campo < nombre del modelo> es el nombre del modelo de la flexión, (@n-os en la figura 3).

Es una cadena de caracteres que no contiene espacios en blanco y que debe comenzar imperativamente por el caracter "@".

El segundo campo <índice de la raíz> es una cifra entre 0 y 9 que indica al compilador el número de caracteres que debe sustraer a la forma lematizada de la palabra que sigue este modelo, para obtener la raíz, en el caso de que ésta no figure explícita en la entrada lexical, (-1 en la figura 3).

El tercer campo <terminación> es una terminación declarada de antemano en el fichero de terminaciones. (+os en la figura 3 para indicar la terminación del tipo *niño*). Si no fuera así, el compilador generaría un error.

El cuarto campo <definición morfológica> es una cadena de caracteres cualquiera, con la salvedad de no contener espacios en blanco ni "@" como primer caracter (n/sm en el ejemplo, para indicar nombre, singular, masculino).

Utilizamos el caracter "*" como variable, para marcar, por ejemplo, la posibilidad de un nombre o adjetivo, de ser masculino o femenino, como es el caso de "responsable" en el ejemplo de la figura 3.

Los modelos de flexión pueden estar repartidos en diferentes ficheros.

2.3 Entradas lexicales:

Una entrada lexical puede definirse de tres formas:

- <L> <modelo de flexión>
<L> es el lema de una palabra².
Es el caso más simple de entrada.

En primer lugar, el programa calcula la raíz del lema sustrayendo los n últimos caracteres; siendo n el índice de la raíz definido en el modelo correspondiente.

Genera, a continuación, las formas flexivas añadiendo a la raíz, las terminaciones declaradas en el modelo; les asocia sus definiciones morfológicas y las organiza en un autómata de estados finitos.

Por ejemplo, declaramos de esta manera los lemas del tipo *niño* (ver figura 4).

2

Seguimos los criterios adoptados por GENELEX (GENELEX, 93) para la elección del lema.

• **<L> <raíz> <modelo de flexión>**

Es el tipo de entrada utilizada para los lemas cuya raíz no puede obtenerse sustrayendo un determinado número de caracteres.

Es el caso del verbo *comunicar* que tiene dos raíces: *comunic-* y *comuniqu-* (ver figura 4).

Así pues, la raíz está explícitamente señalada en el segundo campo de la entrada.

• **<L> <raíz>**

<terminación>₁ <definición morfológica>₁

<terminación>₂ <definición morfológica>₂

.....

<terminación>_n <definición morfológica>_n

En este tipo, la entrada no utiliza un modelo de flexión, ya sea por el escaso número de flexiones, o por que el posible modelo, no declarado, sería poco representativo. Es el caso del artículo, donde "el" es el lema de las raíces "el", "la", "lo", "los", "las" (ver figura 4)

Las entradas pueden declararse en diferentes ficheros.

En la figura 4 pueden apreciarse unos ejemplos de entradas lexicales, suponiendo, claro está, que las terminaciones y los modelos han sido convenientemente declarados (cf. 3.1.1 y 3.1.2).

| # EJEMPLOS DE ENTRADAS LEXICALES | | | |
|----------------------------------|----------|--------|--------|
| ##### | | | |
| animal | @nm-es | | |
| autoridad | @nf-es | | |
| granjero | @n-os | | |
| niño | @n-os | | |
| comunicar | comunic | @v89-1 | |
| comunicar | comuniqu | @v89-2 | |
| el | el | +@ | art/sm |
| el | la | +@ | art/sf |

figura 4

3 Compilación de un diccionario

Todos los nombres de los ficheros del diccionario fuente deben declararse en un fichero llamado "data". La compilación produce un diccionario binario llamado "dict.bin", que será utilizado para el análisis de textos.

Una vez, preparado el fichero *data*, el utilizador puede lanzar la compilación siguiendo las órdenes del menú, sabiendo que si existiese un error de sintaxis o de lógica (por ejemplo la utilización de una terminación que no ha sido declarada) el compilador se parará, indicando un mensaje de error en el que figurarán el tipo de error, el nombre del fichero del diccionario fuente en que se produce tal error, así como la línea donde se localiza.

4 ANÁLISIS DE TEXTOS EN ESPAÑOL

Un análisis morfológico de textos presenta una serie de problemas de tipo lingüístico y tipográfico que se han intentado tener en cuenta para la concepción de Smorph. La experiencia adquirida en el tratamiento de textos con REAL, "Programa de extracción automática de lemas con su contexto" (Rodrigo, 94) (Ait-Mokhtar, Rodrigo y otros 93) nos ha ayudado a repertoriar muchos de estos problemas, que sin un trabajo sobre el terreno, son difíciles de imaginar.

Para la construcción del diccionario fuente del español tenemos presente además de la literatura lingüística (Alarcos, 80) (Hernández, 84) etc., manuales prácticos como, entre otros (EFE, 89) (EL PAÍS, 90); para una bibliografía de base sobre la morfología (Varela, 92).

Para los modelos verbales, seguimos por razones prácticas, (Bescherelle, 88), si bien se podrían reducir el número de modelos allí presentados.

4.1 Prefijos y sufijos

Las nociones de "prefijo" y "sufijo" deben de entenderse en el sentido tipográfico ya que no se corresponden exactamente con la noción tradicional de la lingüística.

Un prefijo es todo ítem que se escribe sin estar separado de la palabra siguiente por un caracter separador (un espacio en blanco, un guión, etc.) (Ait-Mokhtar, 94).

Citemos, como ejemplo el prefijo "ex-" con sus diferentes formas:

ex ex- ex - (excampeón | ex-campeón | ex - campeón)

Otro ejemplo, "para", que puede ser, preposición, verbo o prefijo:

| | |
|--------------|---|
| para | preposición |
| para verbo | parar |
| para prefijo | (paramilitar para-militar para - militar) |

Para distinguir el prefijo "para-" de los otros casos, declaramos para "para" una entrada en la que se especifica que no debe estar seguida de un espacio en blanco.

Sin embargo, es posible encontrar "para militar". En este caso no hay posibilidad de desambiguización en el nivel morfológico.

Los sufijos se tratan de la misma manera que los prefijos.

Los pronombres clíticos son considerados como sufijos, ya que no hay espacio en blanco entre las formas verbales que les preceden y dichos pronombres (Castel, 94).

Cuando los dos clíticos aparecen juntos, la forma resultante es una palabra esdrújula, y por tanto acentuada, lo que exige normalmente una nueva raíz, que se declarará en la entrada lexical como se ve en los ejemplos siguientes:

| | |
|--------|------------|
| -lo | pon-lo |
| -le | di-le |
| -se-lo | pón-se-lo |
| | dió -se-lo |

Los prefijos y sufijos se segmentan como tales. Si no se quiere segmentar una palabra que contiene un prefijo, como *prever*, simplemente basta con declararlo en el diccionario; ya que el analizador da en salida el segmento más largo definido en el diccionario.

Ciertos signos de puntuación, como el punto o la coma, son también considerados como sufijos, dado que no hay ningún espacio en blanco entre el punto y la palabra que le precede.

La declaración de un prefijo o un sufijo se hace con el caracter "^", como se observa en la figura 5.

| | | | | |
|-----|-------|---|---|--------------|
| re- | re^ | + | @ | pref |
| re- | re-^ | + | @ | pref |
| ex- | ex^ | + | @ | pref |
| ex- | ex-^ | + | @ | pref |
| ex- | ex_-^ | + | @ | pref |
| se | ^se | + | @ | pro/**3/refx |
| | | + | @ | pro/**3/obji |
| le | ^le | + | @ | pro/s*3/obji |
| lo | ^lo | + | @ | pro/s*3/objd |
| . | ^. | + | @ | punt/pt |

figura 5

4.2 Palabras compuestas

Una palabra compuesta se define como un segmento formado por varias palabras, que pueden estar separadas por un espacio en blanco, pero que no tienen exigencias sintácticas de orden interno (por ejemplo, exigencias de concordancia entre las palabras componentes).

Por otra parte, ningún elemento externo debe poder intercalarse entre las palabras componentes de una palabra compuesta:

guardia marina
guardia marinas
guardias marinas
acerca de
lo que
tarjeta de crédito

La declaración de una palabra compuesta se hace remplazando los posibles espacios en blanco por el caracter "_" como se observa en la figura 6.

| | | | | |
|----------------|-----------------|---|---|------|
| Guardia_marina | guardia_marina | + | @ | n/s* |
| Guardia_marina | guardias_marina | + | @ | n/p* |
| Guardia_marina | guardia_marina | + | @ | n/p* |
| kilómetro-hora | kilómetro-hora | + | @ | n/sm |
| kilómetro-hora | kilómetros-hora | + | @ | n/pm |

figura 6

4.3 Ambigüedad en la segmentación

Se dan casos de palabras compuestas, y de otros tipos de palabras, que ofrecen una ambigüedad de segmentación que sólo un análisis sintáctico o semántico puede solucionar:

- (1) "la niña se [acerca] de nuevo a su madre"
- (2) "[acerca de] la música popular, el ministro dice..."
- (3) "[a pesar] de todo, Induraáin correrá el Tour"
- (4) "vuelve a [pesar] más de sesenta a cien kilos"

Tenemos dos lecturas posibles de la cadena de caracteres "acerca de". En (1) "acerca" es una forma flexiva del verbo "acercar" y "de" es una preposición, mientras que en (2) "acerca de" es un adverbio.

Igualmente en la cadena "a pesar"; en (3) "a pesar" es un adverbio; mientras que en (4) "a" es una preposición y "pesar" es el infinitivo del verbo *pesar*.

Los puntos finales de las siglas constituyen otro caso de ambigüedad de segmentación, ya que pueden no permitir el reconocimiento del final de una frase:

- (5) la R.E.N.F.E. suprimirá las líneas no rentables
- (6) Habrá huelga en la R.E.N.F.E.

La declaración de un segmento ambiguo se hace con el caracter "~", como se observa en la figura 7.

| | | | |
|-----------|------------|----|------|
| acerca_de | acerca~_de | +@ | prep |
| R.EN.FE. | R.EN.FE~. | +@ | n/sf |

figura 7

Por último, en el caso de los homógrafos, si la declaración hecha en el diccionario es correcta, Smorph dará todas las definiciones morfológicas posibles:

| | | |
|-----------|------|--|
| objetivo | @n-8 | # nombre "objetivo" |
| objetivo | @a-1 | # adjetivo |
| objetivar | @vb5 | # verbo, indicativo presente: "objetivo" |

4.4 Terminaciones distinguidas y palabras desconocidas

Por muy completo que sea un diccionario, el tratamiento automático de textos presenta siempre casos de palabras desconocidas como es el caso de los neologismos, las palabras no reseñadas en el diccionario, los términos técnicos, etc. (Silberztein, 93).

En ciertos casos, a partir de la terminación morfológica podemos, sin embargo, describir la categoría de la palabra desconocida.

Así por ejemplo, cualquiera que sea la cadena de caracteres acabada por "-ción" ("acción", "tracción" "canción"), podemos estar seguros de que dicha palabra es un nombre; e, igualmente, cualquiera que sea la cadena de caracteres acabada por "-ó", podemos asegurar que dicha palabra es una forma flexiva de un verbo en pretérito indefinido.

Smorph marca con el carácter "!" las definiciones morfológicas obtenidas gracias a una terminación distinguida, como se ve en el ejemplo siguiente sacado de un parte de prensa (cf anexo), en el que hay un error tipográfico en la palabra "población":

poiblación < poiblación n/sf >!

4.5 Tolerancia de errores

Es posible tener en cuenta un número de errores tipográficos, que son frecuentes en los textos, declarando como terminaciones las posibles terminaciones erróneas.

Así por ejemplo, en español, podemos considerar ciertos acentos de formas flexivas, sopena de no reconocerlos en los textos por aparecer generalmente no acentuados, como "oír " y "oir" o, aunque menos frecuente, la -a de la cuarta persona del pretérito indefinido, como "vinistes" por "viniste".

El hecho de declarar ciertos signos de puntuación como sufijos y como entradas normales, permite también el reconocimiento de las palabras independientemente de los espacios en blanco.

5 Implementación

El sistema está implementado en C sobre UNIX en una estación SPARC 2. La velocidad de análisis y del etiquetado de textos es de 1000 palabras por segundo aproximadamente.

Es interesante señalar la utilización del espacio de memoria. El espacio utilizado por una entrada puede llegar a 4 octetos por término medio, sobre todo si las formas canónicas tienen una gran cantidad de formas flexivas, como es el caso de los verbos en general y del verbo español en particular.

6 Referencias bibliográficas

(Ait-Mokhtar, Rodrigo y otros, 93)

AÏT-MOKHTAR Salah; DAFNIET Caroline; GANGUTIA Pascal; RODRIGO MATEOS José Lázaro : "REAL: A program for Searching and Extracting Lemmas", DELIS, LRE Project 61-034. SDM-2 Meeting April 3/6.

(Ait-Mokhtar, 94)

AÏT-MOKHTAR Salah : "SMORPH, Segmentation et analyse morphologique; guide d'utilisation", publicación interna, GRIL.

(Alarcos, 80)

ALARCOS LLORACH Emilio: *Estudios de gramática funcional del español*, Madrid, Gredos.

(Bescherelle , 88)

MATEO Francis, ROJO Antonio J. : *Bescherelle. El arte de conjugar en español. Diccionario de 12000 verbos.* Hatier. Paris

(Castel, 94)

CASTEL Victor M : *Paquets de clitiques de l'espagnol. Calcul de la corrélation entre rôles sémantiques et propriétés morphologiques et implémentation dans une grammaire catégorielle d'unification.* Tesis defendida en la universidad de Clermont-Fd.

(EFE, 89)

Agencia EFE, *Manual del español urgente*, Madrid, Cátedra.

(EL PAÍS, 90)

Libro de estilo Madrid, Publicaciones El País.

(GENELEX, 93)

Consortium GENELEX: Rapport Projet Eureka Genelex
"Couche Morphologique".

(Hernández, 84)

HERNÁNDEZ ALONSO César: *Gramática funcional del español*, Madrid, Gredos

(Rodrigo, 94)

RODRIGO MATEOS José Lázaro : "*Programa de extracción automática de palabras con su contexto REAL*". SEPLN nº 14 pags 149 - 163).

(Silberztein ,93)

SILBERZTEIN Max : *Dictionnaires électroniques et analyse automatique de textes.* Paris Masson.

(Varela, 92)

VARELA ORTEGA, Soledad: *Fundamentos de Morfología.*, Madrid, Síntesis.

ANEXO: Análisis morfológico y etiquetado de un texto en español

En este anexo presentamos, un resumen de las abreviaturas utilizadas en las definiciones morfológicas, un texto y el análisis realizado por Smorph.

1 Resumen

| | |
|-------------|--|
| ! | palabra reconocida gracias a una terminación distinguida |
| 3 | 3ª persona |
| a | adjetivo |
| adv | adverbio |
| art | artículo |
| coma | coma |
| conj | conjunción |
| det | determinante |
| guión | guión |
| imp | imperfecto |
| imv | imperativo |
| ind | indicativo |
| indf | indefinido |
| inf | infinitivo |
| MOT_INCONNU | palabra no declarada en el diccionario |
| n | nombre |
| num | numeral |
| par | participio |
| par-a | paréntesis -apertura |
| par-c | paréntesis -cierre |
| pf | plural femenino |
| pi | pretérito indefinido |
| pm | plural masculino |
| pos | posesivo |
| prep | preposición |
| pro | pronombre |
| pt | punto |
| punt | signo de puntuación |
| s* | singular (masculino o femenino) |
| sf | singular femenino |
| sm | singular, masculino, |
| sub | subjuntivo |

2 Texto para analizar

El texto presentado es un parte de prensa del 24 de febrero de 1993.

Un total de 19 personas - 8 policías, 8 cuatrerros y tres granjeros - murieron el lunes pasado en una escaramuza entre las fuerzas del orden y ladrones de ganado en el nordeste de Kenia, anunciaron el miércoles las autoridades.

La policía había detenido en el distrito de Wajir a la banda de cuatrerros responsables del robo de 260 cabezas de ganado. Las tres víctimas civiles eran granjeros llegados a identificar a sus animales.

El sábado último, 27 personas murieron a manos de cuatrerros en el distrito de Pokot (oeste) y un centenar de cabezas de ganado desaparecieron.

En enero, 8 policías y varios paisanos murieron en incidentes con bandidos en Wajir, donde la inseguridad creciente se debe a bandas armadas procedentes de la vecina Somalia, según el gobierno keniano.

Por último, ayer siete personas sospechosas de practicar la hechicería fueron asesinadas por la población en el distrito de Bungoma, en la frontera de Uganda, informó la prensa.

2 Texto analizado por Smorph:

| | | | |
|-------------------|----------------------------|----------------|-----------------------------|
| Un | < un art/sm/indf > | de | < de prep > |
| total | < total n/s* > | 260 | < 260 det/**/num > |
| de | < total a/s* > | cabezas | < cabeza n/pf > |
| 19 | < de prep > | de | < de prep > |
| personas | < 19 det/**/num > | ganado | < ganar v/sm*/par > |
| - | < persona n/pf > | | < ganado a/sm > |
| 8 | < - guión/inciso > | | < ganado n/sm > |
| policías | < 8 det/**/num > | . | < . punt/pt > |
| , | < policía n/pf > | Las | < el art/pf > |
| 8 | < , punt/coma > | tres | < tres det/**/num > |
| cuatreros | < 8 det/**/num > | victimas | < víctima n/pf > |
| | < cuatrero n/pm > | civiles | < civil n/p* > |
| y | < cuatrero a/pm > | | < civil a/p* > |
| tres | < y conj/cop > | eran | < ser v/p*3/ind/imp > |
| granjeros | < tres det/**/num > | granjeros | < granjero n/pm > |
| | < granjero n/pm > | | < granjero a/pm > |
| - | < granjero a/pm > | llegados | < llegar v/pm*/par > |
| murieron | < - guión/inciso > | | < llegado a/pm > |
| el | < morir v/p*3/ind/pi > | a | < a prep > |
| lunes | < el art/sm > | identificar | < identificar v/**/inf > |
| pasado | < lunes n/sm > | a | < a prep > |
| en | < pasar v/sm*/par > | sus | < su pro/s*3/poss > |
| una | < pasado a/sm > | animales | < animal n/pm > |
| escaramuza | < en prep > | . | < . punt/pt > |
| entre | < un art/sf/indf > | El | < el art/sm > |
| las | < escaramuza n/sf > | sábado | < sábado n/sm > |
| fuerzas_del_orden | < entre prep > | último | < último a/sm > |
| y | < el art/pf > | , | < , punt/coma > |
| ladrones | < fuerzas del orden n/pf > | 27 | < 27 det/**/num > |
| de | < y conj/cop > | personas | < persona n/pf > |
| ganado | < ladrón n/pm > | murieron | < morir v/p*3/ind/pi > |
| | < de prep > | a | < a prep > |
| en | < ganar v/sm*/par > | manos | < mano n/pf > |
| el | < ganado a/sm > | de | < de prep > |
| nordeste | < ganado n/sm > | cuatreros | < cuatrero n/pm > |
| de | < en prep > | | < cuatrero a/pm > |
| Kenia | < el art/sm > | en | < en prep > |
| , | < el art/sm > | el | < el art/sm > |
| anunciaron | < nordeste n/sm > | distrito | < distrito n/sm > |
| el | < de prep > | de | < de prep > |
| miércoles | < Kenia np/sf > | Pokot | < Pokot MOT INCONNU > |
| las | < , punt/coma > | (| < (punt/par-a > |
| autoridades | < anunciar v/p*3/ind/pi > | oeste | < oeste n/sm > |
| . | < el art/sm > |) | <) punt/par-c > |
| La | < el art/pf > | y | < y conj/cop > |
| policía | < autoridad n/pm > | un | < un art/sm/indf > |
| había | < . punt/pt > | centenar | < centenar n/sm > |
| detenido | < el art/sf > | de | < de prep > |
| en | < policía n/sf > | cabezas | < cabeza n/pf > |
| el | < haber v/s*1/ind/imp > | de | < de prep > |
| distrito | < haber v/s*3/ind/imp > | ganado | < ganar v/sm*/par > |
| de | < detenido a/sm > | | < ganado a/sm > |
| Wajir | < detener v/sm*/par > | | < ganado n/sm > |
| a | < en prep > | desaparecieron | < desaparecer v/p*3/ind/p > |
| la | < el art/sm > | . | < . punt/pt > |
| banda | < el art/sm > | En | < en prep > |
| de | < distrito n/sm > | enero | < enero n/sm > |
| cuatreros | < de prep > | , | < , punt/coma > |
| responsables | < Wajir MOT INCONNU > | 8 | < 8 det/**/num > |
| del | < a prep > | policías | < policía n/pf > |
| robo | < el art/sf > | y | < y conj/cop > |
| | < banda n/sf > | varios | < varios det/pm/indf > |
| | < de prep > | paisanos | < paisano n/pm > |
| | < cuatrero n/pm > | murieron | < morir v/p*3/ind/pi > |
| | < cuatrero a/pm > | en | < en prep > |
| | < responsable n/p* > | incidentes | < incidente n/pm > |
| | < responsable a/p* > | | < incidente a/p* > |
| | < de el preart/sm > | | |
| | < robo n/sm > | | |

| | | | |
|-------------|----------------------------|-------------|---------------------------|
| con | < con prep > | , | < , punt/coma > |
| bandidos | < bandido n/pm > | ayer | < ayer adv > |
| en | < en prep > | siete | < siete det/**/num > |
| Wajir | < Wajir MOT_INCONNU > | personas | < persona n/pf > |
| , | < , punt/coma > | sospechosas | < sospechoso n/pf > |
| donde | < donde pro/rel > | | < sospechoso a/pf > |
| la | < el art/sf > | de | < de prep > |
| inseguridad | < inseguridad n/sm > | practicar | < practicar v/**/inf > |
| creciente | < creciente a/s* > | la | < el art/sf > |
| se | < se pro/**3/refx > | hechicería | < hechicería n/sf > |
| | < se pro/**3/dat > | fueron | < ser v/p*3/ind/pi > |
| debe | < deber v/s*3/ind/pre > | | < ir v/p*3/ind/pi > |
| | < deber v/s*3/imv > | asesinadas | < asesinar v/pf*/par > |
| a | < a prep > | por | < por prep > |
| bandas | < banda n/pf > | la | < el art/sf > |
| armadas | < armar v/pf*/par > | poiblación | < poiblación n/sf >! |
| | < armado a/pf > | en | < en prep > |
| procedentes | < precedente a/p* > | el | < el art/sm > |
| de | < de prep > | distrito | < distrito n/sm > |
| la | < el art/sf > | de | < de prep > |
| vecina | < vecino n/sf > | Bungoma | < Bungoma MOT_INCONNU > |
| Somalia | < vecino a/sf > | | < , punt/coma > |
| , | < Somalia np/sf > | en | < en prep > |
| según | < , punt/coma > | la | < el art/sf > |
| el | < según prep > | frontera | < frontera n/sf > |
| gobierno | < el art/sm > | de | < de prep > |
| | < gobierno n/sm > | Uganda | < Uganda np/sf > |
| keniano | < gobernar v/s*1/ind/pre > | | < , punt/coma > |
| | < keniano a/sm > | informó | < informar v/s*3/ind/pi > |
| | < keniano n/sm > | la | < el art/sf > |
| , | < . punt/pt > | prensa | < prensa n/sf > |
| Por_último | < por_último adv > | . | < . punt/pt > |